



Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim



Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction

Beatriz López^a, Ferran Torrent-Fontbona^{a,*}, Ramón Viñas^a,
José Manuel Fernández-Real^{b,c}

^a University of Girona, Campus Montilivi, building EPS4, 17071 Girona, Spain

^b Biomedical Research Institute of Girona, Avda. de França, s/n, 17007 Girona, Spain

^c CIBERObn Pathophysiology of Obesity and Nutrition, Instituto de Salud Carlos III, Madrid, Spain

ARTICLE INFO

Article history:

Received 10 February 2017

Received in revised form 4 September 2017

Keywords:

Type 2 diabetes

Random Forest

Feature learning

Predictive model

Gini importance

ABSTRACT

Objective: The use of artificial intelligence techniques to find out which Single Nucleotide Polymorphisms (SNPs) promote the development of a disease is one of the features of medical research, as such techniques may potentially aid early diagnosis and help in the prescription of preventive measures. In particular, the aim is to help physicians to identify the relevant SNPs related to Type 2 diabetes, and to build a decision-support tool for risk prediction.

Methods: We use the Random Forest (RF) technique in order to search for the most important attributes (SNPs) related to diabetes, giving a weight (degree of importance), ranging between 0 and 1, to each attribute. Support Vector Machines and Logistic Regression have also been used since they are two other machine learning techniques that are well-established in the health community. Their performance has been compared to that achieved by RF. Furthermore, the relevance of the attributes obtained through the use of RF has then been used to perform predictions with *k*-Nearest Neighbour method weighting attributes in the similarity measure according to the relevance of the attributes with RF.

Results: Testing is performed on a set of 677 subjects. RF is able to handle the complexity of features' interactions, overfitting, and unknown attribute values, providing the SNPs' relevance with an up to 0.89 area under the ROC curve in terms of risk prediction. RF outperforms all the other tested machine learning techniques in terms of prediction accuracy, and in terms of the stability of the estimated relevance of the attributes.

Conclusions: The Random Forest is a useful method for learning predictive models and the relevance of SNPs without any underlying assumption.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

There is a great deal of interest in finding the SNPs that are related to a given illness in order to appropriately develop a corresponding personalized treatment. The first approaches with regard to studying the relationships between SNPs and diseases focused on single individual variable analysis, where a variable (SNP) is removed, and then some predictor indicator is analysed to measure the impact of the variable influence. However, interactions between variables meant that this approach did not perform well.

Therefore, other approaches based on machine learning techniques, which enable the analysis of multiple combinations of variables, are preferred [1,2].

The authors in [3] provide an overview of the different machine learning techniques applied to SNP data, from which two main approaches are distinguished: SNP association studies and predictive modelling. While SNP association studies consists of grouping SNPs according to their expression profiles (e.g. molecular function, biological process, cellular components), predictive modelling aims to identify which features are relevant to a specific function or class. For example, which features are particularly relevant with regard to Type 2 diabetes (T2D).

Concerning the use of predictive models with SNP data, these methods suffer from the dimensionality problem: hundreds of subjects (samples), with thousands of SNPs per subject (features, attributes). As a consequence, [3] warns about the risk of incur-

* Corresponding author.

E-mail addresses: beatriz.lopez@udg.edu (B. López), ferran.torrent@udg.edu (F. Torrent-Fontbona), rvinast@gmail.com (R. Viñas), jmfreal@idibgi.org (J.M. Fernández-Real).

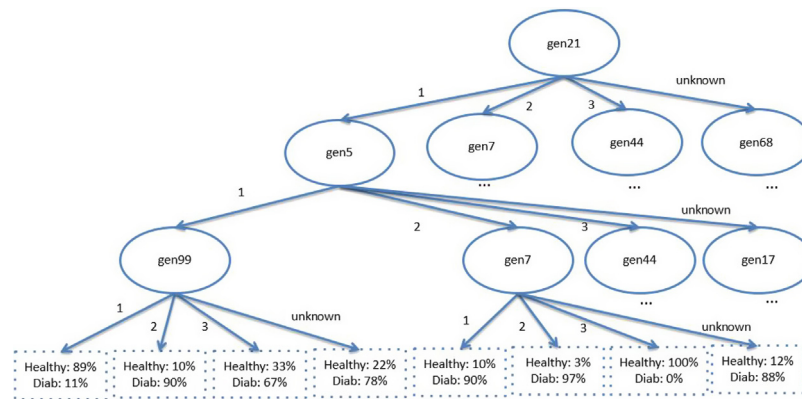


Fig. 1. Example of a decision tree.

ring an overfitting problem [4] when applying machine learning techniques to such kinds of datasets.

To overcome the overfitting problem, regularization techniques [4] are applied, but they present some difficulties regarding tuning the appropriate regularization parameter [5] and moving away from the current trends of personalized medicine [6]. On the other hand, the Random Forest (RF) technique [7] has been proven to outperform most of the current machine learning techniques when it comes to building classification models in general, and predictive models in particular, without any underlying assumptions [8]. Moreover, it is a computationally efficient technique and one that is almost free of parameters. The RF technique consists of building a given number of decision trees (a forest), which are combined in an ensemble mechanism (e.g. majority voting) in order to obtain a final classification outcome (e.g. ill or healthy), with a confidence degree associated with the result (a prediction indicator).

The relevance of features that conform to the RF model is obtained by aggregating the relative importance of the features over all of the trees [9,10]. Therefore, no particular pre-processing techniques are required for feature selection [11,12]. Moreover, as an ensemble technique, the obtained relevance feature set is stable [13,14].

An additional property of the RF technique is its capacity for handling missing information [15], a common situation when dealing with SNP data [16]. This is due to the ensemble nature of the RF method, which combines several decision trees to provide a classification outcome (e.g. prediction) [7]. Each decision tree is learned by using a subset of features (SNPs) that are randomly selected, as well as a subset of samples that are also randomly chosen. However, the RF technique does not remove any information, maintaining the changes towards a personalized outcome.

This paper addresses the application of the RF technique to a dataset of SNPs, which has a significant percentage of missing information, classified in terms of people with T2D and people without it. In particular, our work concerns the identification of the relevant SNPs related to T2D. Furthermore, RF performance is compared with other well-established machine learning techniques, such as Support Vector Machines (SVM) and Logistic Regression (LR).

2. Material and methods

The Biomedical Research Institute of Girona has been gathering information about the SNPs of subjects with their corresponding diagnoses (T2D, glucose intolerance), as well as that of healthy subjects. Based on the available data, a T2D risk prediction model has been obtained with the use of the RF technique, from which the relevance of each feature is obtained by using the Gini importance [9].

2.1. The problem

The problem addressed in this paper is to find the relevance of a set of SNPs g_1, g_2, \dots, g_n , given a set of samples P corresponding to people with and without T2D, in order to enable the prediction of T2D.

Each sample is noted as (x, y) , where x is a list of attribute-value pairs $\langle g_i, v_i \rangle$ regarding the SNPs g_1, g_2, \dots, g_n and their values v_1, v_2, \dots, v_n for the given sample; and y is the class to which the person belongs. In this particular case, $y \in C = \{\text{healthy}, \text{Type2diabetes}\}$. Attributes, SNPs and features are used synonymously throughout this work.¹

Each SNP i has NVA_i values. In our particular case, $NVA_i = 4$ ($\forall i$), with the following interpretation: (1) the SNP is not present; (2) the SNP is present; (3) the SNP has been expressed; (4) *unknown* value. Therefore, we are considering SNPs with missing information.²

2.2. Random Forest

RF is a supervised learning method, which means that each instance or sample is labelled with the outcome (class).

RF consists of an ensemble of k classifiers $h_1(x), h_2(x), \dots, h_k(x)$, with $h(x)$ being the joint classifier [7,18]. Each classifier $h_i(x)$ consists of a decision tree, in which nodes are attributes (see Fig. 1). The selection of which attribute is collocated in a node n is performed as follows: (1) a subset of attributes is randomly selected, (2) an evaluation measure is applied to the selected attributes according to their capability for providing homogeneity partitions of the samples, and (3) the attribute with the highest score is chosen. In particular, we use the change of the Gini impurity³ [18] to compute the score, as described in Eq. (1)

$$\Delta G(g_i, n) = - \sum_{C_k \in C} p^2(C_k) + \sum_{j=1}^{NVA_i} p(v_{i,j}) \sum_{C_k \in C} p^2(C_k | v_{i,j}) \quad (1)$$

where $v_{i,j}$ is the j value of the i SNP. Probabilities are estimated according to the instances that reach the n node.

Once a node is assigned to an attribute g_i , the data is split into as many sets as values the g_i attribute has (four). Then, the tree is grown with new nodes in each branch. These are obtained by

¹ Attributes is often the proper notation of supervised machine learning methods; SNPs of genetics, and features of feature learning methods.

² In fact, this could be considered as a unique-value imputation method, as the unknown or missing value is treated as another attribute value [17].

³ Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

Download English Version:

<https://daneshyari.com/en/article/6853331>

Download Persian Version:

<https://daneshyari.com/article/6853331>

[Daneshyari.com](https://daneshyari.com)